Pers Ubiquit Comput
DOI 10.1007/s00779-007-0172-1

ORIGINAL ARTICLE

Integration of audiovisual sensors and technologies in a smart room

Joachim Neumann · Josep R. Casas · Dušan Macho · Javier Ruiz Hidalgo

Received: 1 August 2006 / Accepted: 15 December 2006

Abstract At the Technical University of Catalonia (UPC), a smart room has been equipped with 85 microphones and 8 cameras. This paper describes the setup of the sensors, gives an overview of the underlying hardware and software infrastructure and indicates possibilities for high- and low-level multi-modal interaction. An example of usage of the information collected from the distributed sensor network is explained in detail: the system supports a group of students that have to solve a lab assignment related problem.

1 Introduction and motivation

The smart room at UPC has been designed to hold group meetings, presentations and undergraduate courses in small groups. The room serves two purposes: first, it is an experimentation environment, where researchers test multimodal analysis and synthesis developments in the area of human–computer interfaces; second, it doubles as a data collection facility for research purposes, providing data for technology development and evaluation. To this end, the room has been setup with audio–visual sensors and computing equipment. The multimodal integration of the sensors in a distributed sensor network aims at providing services to the participants in the smart room. The software architecture that allows handling the high-bandwidth data streams is based on distributed computing and allows going beyond the computing capabilities of non-integrated computer and sensor-networks.

The UPC smart room permits implementation and testing of a large variety of audio technologies, such as Automatic Speech Recognition, Speaker Identification, Speech Activity Detection, Acoustic Source Localization, Acoustic Event Detection and Speech Synthesis.

For video technologies, the multicamera setup in the smart room allows experimenting with visual analysis technologies that strongly rely on exploiting the available redundancy when the same scene is seen from up to eight different cameras. Not only 3D visual analysis is possible in the smart room, but also any 2D visual analysis approach can be improved by selecting at any time the best camera for a given analysis task. The list of video technologies currently being developed in the smart room are Multi-Camera Localisation and Tracking, Face Detection, Face ID, Body Analysis and Head Pose Estimation, Gesture Recognition, Object Detection and Analysis, Text Detection and Global Activity Detection

In addition, multi-modal approaches (audio + video) are being currently investigated for the Person Identification and Person Localization and Tracking technologies.

A specific application demonstrating the capabilities provided by these technologies has been implemented at UPC. The resulting service provided to the end user is

This work has been partially supported by the European Union, IP 506909 (CHIL)

J. Neumann (✉) · J. R. Casas · D. Macho · J. R. Hidalgo
Signal Theory and Communications Department, UPC—Technical University of Catalonia, Campus Nord edifici D5 Jordi Girona 1-3, 08034 Barcelona, Spain
e-mail: joachim@gps.tsc.upc.edu

J. R. Casas
e-mail: josep@gps.tsc.upc.edu

D. Macho
e-mail: dusan@gps.tsc.upc.edu

J. R. Hidalgo
e-mail: jrh@gps.tsc.upc.edu

called ''Memory Jog'', because it helps participants in the smart room by providing background information and memory assistance. The Memory Jog is provided in an educational scenario, in which a group of students is asked to solve a problem related to a leaning tower (like the one in Pisa). Several services are provided to both the students (who are situated in the UPC smart room) as well as to the teacher who can be either in the smart room or at his office.

In this paper, we describe the sensor setup in the smart room, the applied analysis technologies developed, the software architecture as the basis for the integration of these technologies into the actual service provided by the room and, finally, the Memory Jog as an instantiation of context aware service in a particular environment.

2 Sensor setup

In order to provide the functionalities of the Memory Jog service to the group of students, the distributed sensor network needs to identify the participants in the room, track their positions over time as well as detect speech and identify voices. The system is capable of continuous monitoring the of UPC smart room [1]. It provides the necessary infrastructure to perform an audio–visual scene analysis as well as a basic modeling of room scenarios.

2.1 Audio sensors

The multi-microphone network provides audio data for analysis of the acoustic scene in the smart-room by employing several audio technologies such as detection and localization of multiple acoustic events, speech activity detection and speech recognition, speaker localization and tracking, etc.

Several kinds of audio sensors are installed in UPC smart-room. A NIST 64 microphone array Mark III [2] provides a high-resolution audio signal (44.1 kHz, 24 bit) with all 64 channels synchronized by a word clock. The array is placed close to the wall approximately 4 m from the main talker area (see Fig. 1). Three T shaped microphone clusters consisting of four microphones are positioned on three walls except the wall with Mark III (see Fig. 1) at a height of about 2 m. Similarly to Mark III, the clusters provide word clock synchronized high-resolution audio signals. Moreover, four omni-directional microphones are placed on the table without having a fixed position, and five barely visible close-talking microphones (Countryman) can be attached to the meeting participants; their signal is wirelessly transferred allowing the participants to move freely.

The mentioned audio sensors can be used for various tasks, but some sensors suit better certain tasks than the others. For example, the Mark III is mostly used for Automatic Speech Recognition of the beam-formed signal, but it can also be used for Audio Source Localization. Similarly, the T-shaped clusters are mostly used for Audio Source Localization, but they may also be used for Automatic Speech Recognition, if speaker is near a specific cluster.

2.2 Video sensors

The four cameras placed in the room corners aim at covering the whole area of the smart room. Their angle and position ensures that each object of interest in the room is simultaneously covered by at least two cameras. These cameras are used for overall monitoring of the room, detecting and tracking people, body modeling and classification of activities [3]. These corner cameras are also used for Face ID, Head Pose Estimation and Gesture Recognition. The fifth camera is positioned at the ceiling of the room. This zenithal camera provides valuable help for Person Localization and Tracking and Global Activity Detection [4]. Cameras six and seven are positioned on the long walls and point at participants seated at the table [5, 6]. They are mostly intended for Face ID, Head Pose Estimation and Gesture Recognition. The eighth camera is an active pan–tilt–zoom camera that can be pointed at the person of interest, e.g., the presenter. Initially it points at the door to capture a high-resolution image of the face of people entering the smart room which served for Face Detection and Face ID.

3 Applied analysis technologies

In the following, the audio and video technologies being developed at UPC are listed:

- Multi-Camera Localization and Tracking: This technology consists of two steps: Firstly, in each of the five cameras, regions of interest (e.g., persons, chairs or laptops) are detected via foreground segmentation. The result of this step is five binary foreground masks. In the second step, a three-dimensional representation of the regions of interest is obtained by a ''ShapeFromSilhouette'' algorithm [7, 8] that receives the binary foreground masks from all five cameras. In this step, these three-dimensional regions of interest are labeled and tracked over time. As side effect of the three-dimensional analysis, the robustness and consistency in the original 2D FG regions can be improved by

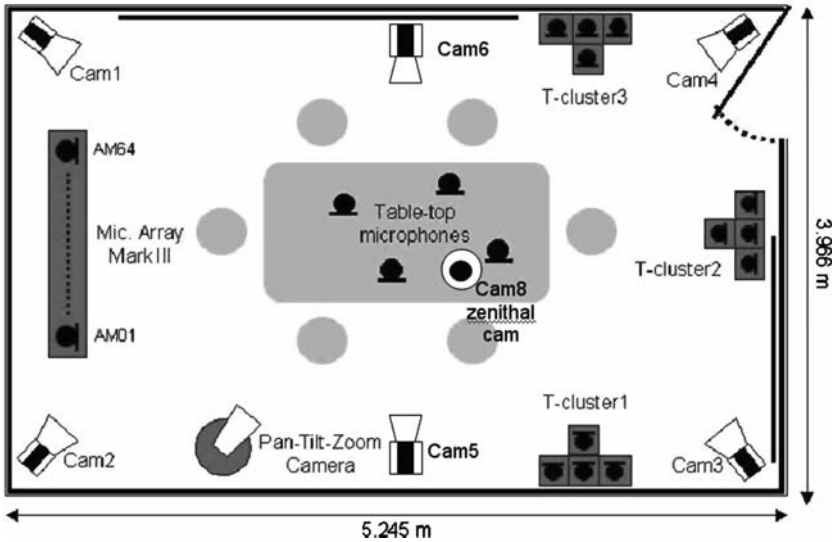

Fig. 1 Sensor set-up of the smart room at UPC: the multi-sensor system consists of various audio and video sensors

re-projecting the three-dimensional regions of interest upon the two-dimensional binary foreground masks.

- Blob Analysis: This technology further analyses the three-dimensional regions of interest in order to distinguish objects such as chairs from people. This technology also analyses human body posture (standing, sitting, etc.) and detects gestures such as a raised arm. To achieve this, a standard model of the human body is aligned to the three-dimensional regions of interest earlier classified as 'person'. The parameters of the joints and nodes of the human model are updated over time to yield a real-time representation of the person (considering the restrictions of the human body model). These parameters are used for gesture recognition.
- Face detector: This technology detects faces and creates a mask of the part of the image that contains the face. The face detection is only applied on those parts of the image that have previously been classified as two-dimensional regions of interest (binary foreground masks). The output of this analysis is a binary face mask.
- Face ID: In this technology, several binary face masks corresponding to the output of the previous technology at different time instants are analyzed to select a frontal view of the face. The frontal view is matched against faces stored in a database of faces. In this database, faces from the people that potentially enter the smart room are stored. If no frontal view is available, the algorithm is capable to base the Face ID on side and profile views, although the identification is less reliable. The ID of the face can easily be assigned to the corresponding three-dimensional region of interest to enrich the output of the Multi-Camera Localization and Tracking technology with the Face ID information.
- Object Detector: In this technology, the three-dimensional regions of interest that have been classified as objects are further analyzed. A model-based classification algorithm is used to detect objects of some predefined classes. In the case of a laptop, for example, their state can be further analyzed: lid open versus closed and laptop on versus laptop off. The result of this analysis can be combined with the three-dimensional position of the people in the room to detect if someone is using the laptop. In this case, the screen of the laptop can be potentially used by the Sensor Network as an output device to communicate with the people in the room.
- RoomStatus: This technology is based on a simple foreground pixel-counter that detects activity in predefined areas of the room (e.g., door open/door closed, activity around the coffee-table, etc.) based on a simple threshold criteria with hysteresis.
- Speech Activity Detection: This audio technology provides information about speech activity in the room. The UPC Speech Activity Detection system [9] is based on Linear Discriminant Analysis features extracted from Frequency Filtering parameters; a Decision Tree is used as classifier. Currently, the output of SAD is binary: Speech or Non-speech. Due to low complexity, nearly a hundred Speech Activity Detection systems can be running simultaneously in real-time on the smart-room computer hardware.
- Speaker Identification: This technology provides information about the identity of the active speaker. The UPC Speaker Identification system is based on Gaussian mixture modeling. As acoustic features, we use Mel-frequency Cepstral coefficients and Frequency Filtering features. One Gaussian mixture model is trained for each speaker and during the testing or

on-line running, the model that best matches the incoming signal is selected.

- Acoustic Localizator: This technology offers at each timestamp a three-dimensional position of the active acoustic source or several sources; it can be a speaking person, but also a ringing phone or moving chair. The UPC Speaker Localization and Tracking system is based on the cross-power spectrum phase approach [10], which we showed is quite robust to the speaker head orientation [11] if using an appropriate distribution of microphone arrays. We use three T-shaped microphone arrays and Mark III, so that there is a microphone array at each wall in the room. The system runs in real-time.
- Acoustic Event Detection: The objective of this technology is to detect and classify various acoustic events that may occur in a smart-room, such as door opening/closing, phone ringing, chair moving, and also vocal tract produced non-speech sounds such as cough, laugh, etc. Acoustic Event Detection is a relatively new area and at UPC we currently focus on the investigation of appropriate features and classification/detection methods. In our publications [12–14], we compare and combine ASR features and acoustic features. Also, we showed that the support vector machine approach provides a good classifier alternative to the more common approaches such as Gaussian mixture models.
- Speech Synthesis: This technology is used as the 'Voice of the Room' to address the people in the room. It can either synthesize speech or play a pre-recorded message. The sound output is driven by a politeness-module that monitors the output of the speech activity detector to avoid interrupting human–human communication.
- Answer: This technology provides a Question and Answering engine. It generates the answer from a database that can be adapted to the various domains. This technology has been provided by the Natural language processing Group at UPC. However, since this paper is on multimodal integration, we do not go into details here.

While the technologies listed above are generic, the following technologies have been designed specifically for the Memory Jog Service implemented at UPC.

- Highlight Recorder: This technology serves to record important events. All events detected by any of the abovementioned technologies are analyzed here. Events classified as important trigger the recording of a Highlight describing this event. Each Highlight consists of a text describing the event (e.g., ''People enter the room'' or ''The teacher responds to a question''). This text is superimposed on a snapshot from the camera that provides the optimal view on the event (typically the camera furthest away from the location of the event). These Highlights are used in two technologies explained below: the Student Monitoring System and the Proactive Help module.
- Student Monitoring System: This technology consists of a graphical user interface that allows the teacher to monitor the activity in the room by browsing through the recorded highlights. The teacher can decide to give hints to the students if their progress is slow. The Student Monitoring System is also capable to inform the teacher when one of the students calls the teacher by raising his arm—as detected by the Blob Analysis technology. The teacher can acknowledge the student's call (''I am coming down'') or ignore the request. In the latter case, the Student Monitoring System would initiate the Proactive Help.
- Proactive Help: This (simple) technology displays previously recorded Highlights to the students. These Highlights can serve the students as a hint to progress in their work. Typically, the Highlights display commented work done by a previous group.
- Student Service and QA Interface (cf. Fig. 6): This technology allows the students to select specific details of the assignment and to interact with the Answer technology that provides factual information needed to solve the problem. The Student Service and QA Interface also monitor the progress of the students by following the dialogue between the students and the Answer technology. Furthermore, it registers the final solution provided by the students.

4 Software architecture

The technologies described in the previous section generate a high-bandwidth data stream of several hundreds of Megabytes per second and their results need to be collected in a central logic. The software architecture chosen in the UPC smart room is based on NIST smartflow system [15] and Chilix [16] (XML messaging system developed within the European project IP 506909 CHIL). This is illustrated in Fig. 2.

The lower level of the software architecture consists of the video and audio sensors. These are implemented as smartflow clients in the computers with the corresponding acquisition hardware. The resulting data streams are transferred as smartflow flows into other computers that can either (a) pre-process this data streams (e.g., the foreground segmentation which is part of the Multi-camera Localization and Tracking) in order to provide pre-processed data streams to other technologies or (b) directly

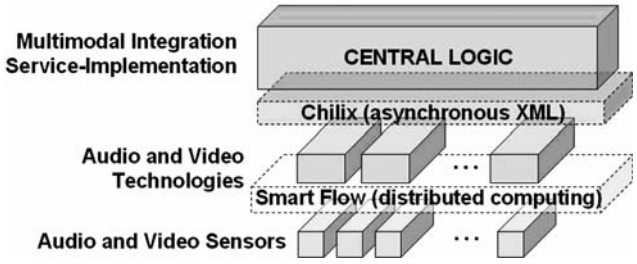

Fig. 2 Software architecture in the UPC smart room

analyze the raw data streams (e.g., Speech Activity Detection audio technology).

Smartflow also provides a mechanism to manually configure a distributed computing system optimized for the task on hand. This consists of the flexibility to decide on which computer in the smart room network a specific technology should run while assuring the correct handling of the data streams between the involved technologies. An example of such a distributed computer processing on the smart room network with several processing modules implementing technologies and high bandwidth data streams is given in the following diagram:

Chilix provides means to easily collect the asynchronous results of the data analysis of these technologies into a central logic that allows high-level multimodal integration. The output of the smartflow clients which implement the abovementioned audio and video technologies is fed asynchronously as XML messages into the common central logic framework. (Fig. 3)

The central logic framework is responsible for visualizing and combining these analysis results. In some cases, a high-level multimodal integration can be realized here. For example, both the results of the Face Detector and the Speech Activity Detection technologies are used to detect that someone enters the smart room. The central logic framework implements a state-model that is adapted to the service to be provided. Changes in the state-model are triggered by events detected in a scene analysis which is based on all video and audio technologies. The following Fig. 4 presents the graphical user interface (GUI) of the central logic for the case of the UPC Memory Jog. (Fig. 5)

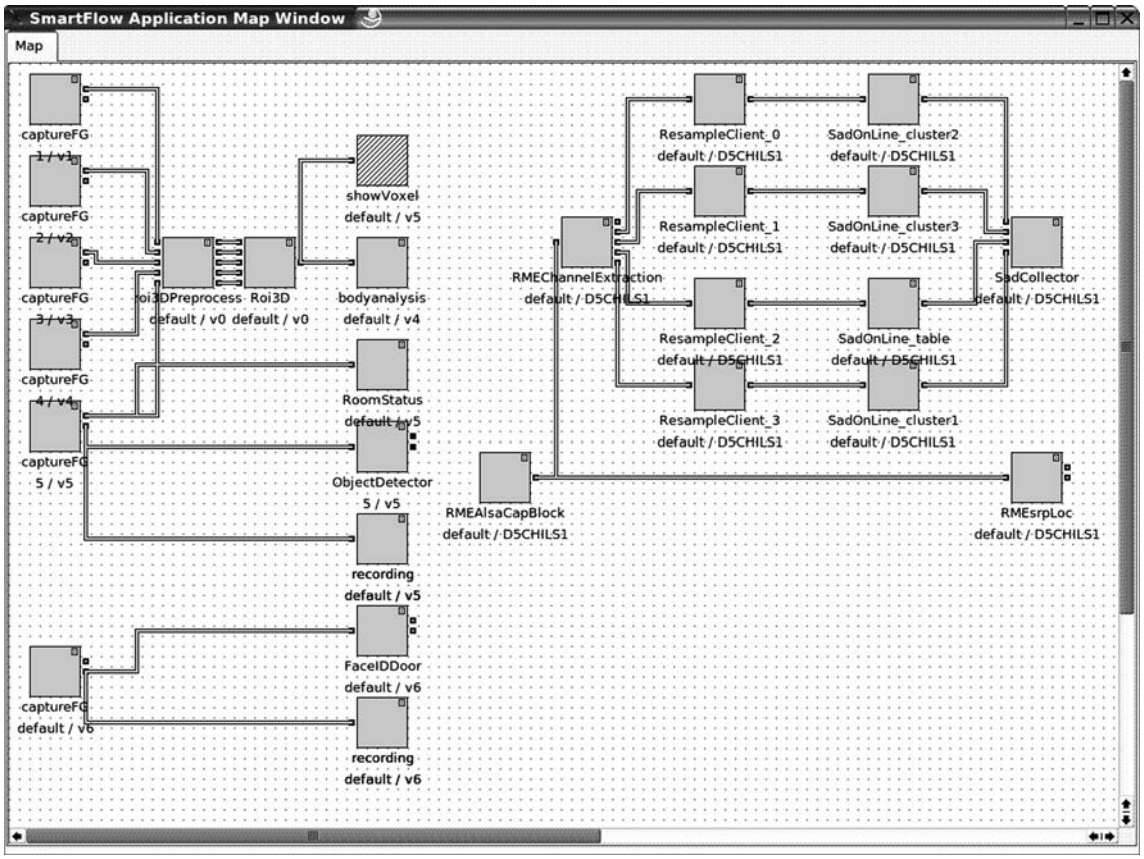

Fig. 3 This diagram shows several technologies implemented as smartflow clients (*squares*) running on six computers in the smart room network, and the data flows between them (*lines*). Some of the technologies as described above consist of several smartflow clients (for example, the multi-camera localization and tracking technology consists of the two smartflow clients ''roi3DPreprocess'' and ''Roi3D'')

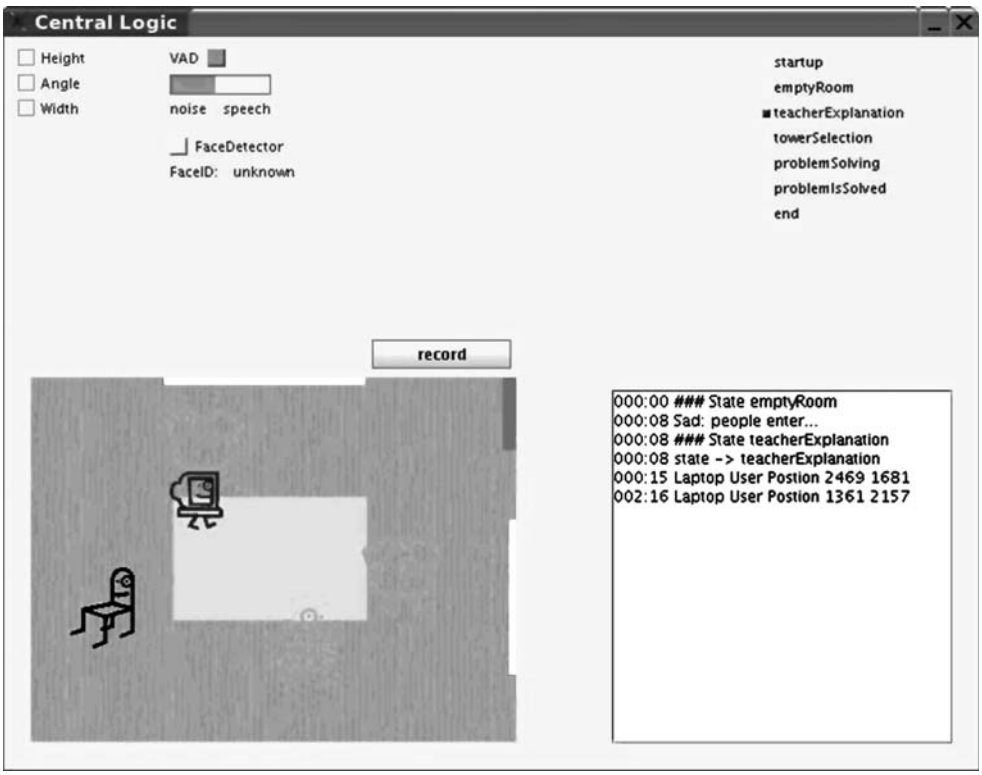

Fig. 4 GUI of the central logic software that receives the results of several video and audio technologies. These results are illustrated in the *left side* of the GUI of the central logic. For instance, the detected position of the chairs, laptops and people in the smart room is indicated in the *lower-left picture*. On the *right side*, the present state in the state-model is indicated and the events occurred in the smart room are listed

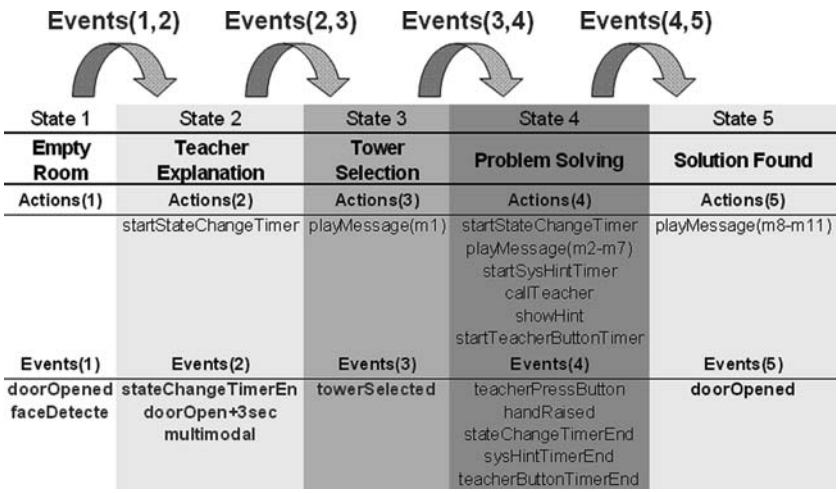

Fig. 5 State model implemented in the central logic

5 The Memory Jog service

The basic concept of the Memory Jog is based on information shift in time and space. For example, the Memory Jog seeks, finds and retrieves information on demand using the Question and Answering technology. The Memory Jog can also translate requests/notice/advice (information shift) from a different instant in time, from databases located elsewhere on the internet, or from one user to another user. This information can be provided reactively (on request) or proactively (automatically). In the latter case, it has to be provided as a help without diverting the user from his main task (unobtrusively). To achieve this, the service needs to be context-aware. The system analyzes what is happening in the smart environment to determine the most appropriate instant in time to interrupt for providing information. The implementation of the service further strives at providing the correct information in a polite manner. Table 1) illustrates the events which trigger the Memory Jog service in the educational scenario where the students solve the leaning tower problem:

Table 1 List of events that are noticed by the system, an explanation how the system has learned about the event and the action taken by the system

Event	How detected	Action
People enter the room	Multimodal detection	Starts perception and analysis
Teacher selects and explains the problem	Interaction with the service interface	Stores information in database, initialized internal timers (for Q&A)
Teacher leaves	Multimodal detection	Starts interacting with students through its voice (a pre-recorded message is played)
A questions is asked	Interaction with the service interface	The system answers and notes down if a relevant information has been requested
An assumption about the task is made	Interaction with the service interface	The system notes down the assumption
Someone raises his hand	Video technologies	The system calls teacher through the teacher GUI
The teacher responds to the student's request or he does not respond to the student's request	Timeouts in the student monitoring system	The system informs students about the teacher's arrival or it gives a pre-recorded hint.
The progress of the students is slow	Timer	The system sends a pre-recorded hint. This service is proactive.
The students have reached a solution	Interaction with service interface	The system notes down
The students are leaving the room	Multimodal detection	The system plays a goodbye message and gives further instructions

This knowledge is fed into a state model in the central logic in order to understand what is going on in the room. Knowledge about the status of the situation model and the collected data allow services to be provided to the participants of a meeting taking place in the smart room. The state model contains information about:

- The state of the meeting: empty room, teacher explanation, tower selection, problem solving, solution found.
- The state of each participant: ID information, speaking/non-speaking, position changes, gestures.
- The state of objects: location and classification of objects on the table.
- Acoustic events.

The services provided by the UPC Memory Jog aims both at the students and the teacher. The services provided to the students are

1. Memory Jog provides information through Q&A (as illustrated in Fig. 6).
2. Students can call the teacher raising their hand.
3. Memory Jog provides proactive help if students don't progress.
4. The voice of the service is polite (does not interrupt).

The services provided to the teacher are:

1. Teacher can supervise the lab session from his office.
2. Teacher gets notice of requests from students and may react.
3. If the teacher does not respond, Memory Jog provides hint.

6 Conclusion

Context awareness is the key requirement to implement the Memory Jog service in the UPC smart room. In order to gather the required information for the context awareness, all of the audio and video technologies mentioned in this paper need to run simultaneously and distributedly in the smart room computer network. The chosen software architecture proved suitable to exchange high-bandwidth data streams between the corresponding computers and to supply asynchronous analysis results to the central logic framework.

The audio and video technologies presented in this paper provide the information that is required to update the state-model and context awareness in the central logic framework. Based on this high-level multi-modal analysis results, a computer-based system could be implemented that interacts with humans in the smart room and provides re-active and pro-active services to them.

The implemented Memory Jog provides useful information to the students and the teacher in and outside the smart room. The quality of the context awareness achieved by the Memory Jog allowed us to even implement services that pro-actively interact with humans.

Integration of audiovisual sensors and multimodal technology modules for analysis and synthesis would not have been possible without the described software architecture (smartflow, Chilix and the central logic). The Memory Jog service running at UPC is just an instantiation of a context aware service. The Memory Jog service demonstration is intended to prove the usability in real

Fig. 6 Graphical user interface of the Student Service and QA Interface

world situations of state of the art multimodal interface technologies mediating in human–computer interaction, and providing help in human and collaborative working environments.

References

1. Josep R, Casas R, Stiefelhagen et al (2004) Multi-camera/multi-microphone system design for continuous room monitoring, CHIL-WP4-D4.1-V2.1-2004-07-08-CO, CHIL Consortium Deliverable D4.1
2. Stanford V, Rochet C (2003) The NIST Mk-III microphone array, and applications of adaptive beam forming to speech. In: 5th international workshop on microphone array systems: theory and practice. (http://www.nist.gov/smartspace/cmaiii.html)
3. Landabaso JL, Xu LO, Pardas M (2004) Robust tracking and object classification towards automated video surveillance. In: Proceedings of the international conference on image analysis and recognition ICIAR 2004, Porto, Portugal, September 29–October 1 2004, Part II, pp 463–470
4. Landabaso JL, Pardàs M, Xu LQ (2005) Hierarchical representation of scenes using activity information. In: Proceedings of ICASSP, Philadelphia, 18–23 March 2005
5. Josep R, Casas O, Garcia, et al (2004) Initial multi-sensor selection strategy to get the best camera/microphone at any time, CHIL-WP4-D4.2-V2.0-2004-10-18-CO, CHIL Deliverable D4.2, October
6. Garcia O, Casas JR (2005) Functionalities for mapping 2D images and 3D world objects in a Multicamera Environment. In: 6th international workshop on image analysis for multimedia interactive services (WIAMIS), Montreux, Switzerland
7. Laurentini A (1994) The visual hull concept for silhouette-based image understanding. IEEE Trans Pattern Anal Mach Intell 16(2):150–162
8. Landabaso JL, Pardas M Foreground regions extraction and characterization towards real-time object tracking. In: Proceedings of joint workshop on multimodal interaction and related machine learning algorithms (MLMI '05), 2005. 3
9. Padrell J, Macho D, Nadeu C (2005) Robust speech activity detection using LDA applied to FF parameters. In: Proceedings of ICASSP'05, Philadelphia
10. Omologo M, Svaizer P (1994) Acoustic event localization using a crosspower-spectrum phase based technique. In: Proceedings of ICASSP'94, Adelaide
11. Abad A, Macho D, Segura C, Hernando J, Nadeu C (2005) Effect of head orientation on the speaker localization performance in smart-room environment. In: Proceedings of INTERSPEECH–EUROSPEECH 2005, Lisbon
12. Temko A, Macho D, Nadeu C (2005) Selection of features and combination of classifiers using a fuzzy approach for acoustic event classification. In: Proceedings of the 9th European Conference on speech communication and technology, Interspeech 2005, Lisbon

13. Temko A, Macho D, Nadeu C (2006) Improving the performance of acoustic event classification by selecting and combining information sources using the fuzzy integral. Lecture notes in computer science (LNCS), vol 3869
14. Temko A, Nadeu C (2006) Classification of Acoustic events using SVM-based clustering schemes. Pattern recognition. Elsevier, Amsterdam (in press)
15. NIST smart space system. http://www.nist.gov/smartspace
16. CHIL Deliverable D2.4 (2006) CHIL software architecture version 2.0